PRJ python

* Thực hiện các cơ chế CRUD (Create – Read – Update – Delete) trên tập dữ liệu.
* Làm sạch dữ liệu
  + **1. Xử lý giá trị thiếu (Missing Values)**
  + **Loại bỏ các dòng hoặc cột** chứa quá nhiều giá trị bị thiếu.
  + **Thay thế các giá trị thiếu** bằng trung bình (mean), trung vị (median), giá trị phổ biến nhất (mode) hoặc giá trị dự đoán từ các thuật toán khác.
  + **Giữ nguyên giá trị thiếu** nếu nó có ý nghĩa (ví dụ: trường hợp "không biết" trong dữ liệu khảo sát).
  + **2. Loại bỏ hoặc xử lý dữ liệu trùng lặp (Duplicates)**
  + Tìm và loại bỏ các dòng dữ liệu bị trùng lặp. Điều này thường xảy ra do việc nhập dữ liệu nhiều lần.
  + **3. Chuẩn hóa dữ liệu (Standardization/Normalization)**
  + **Định dạng lại** dữ liệu sao cho nhất quán (ví dụ: chuyển tất cả chữ thành chữ thường hoặc thống nhất kiểu định dạng ngày tháng).
  + **Chuẩn hóa đơn vị đo lường** để đảm bảo dữ liệu có thể so sánh được (ví dụ: đồng nhất đơn vị tiền tệ, khối lượng).
  + **4. Xử lý dữ liệu không hợp lệ (Invalid Data)**
  + Kiểm tra các giá trị không hợp lệ hoặc không nằm trong phạm vi mong đợi. Ví dụ: dữ liệu nhiệt độ không thể nhỏ hơn -273°C hoặc tuổi không thể lớn hơn 150.
  + **Loại bỏ** hoặc sửa đổi các giá trị không hợp lệ bằng cách xác định nguyên nhân và điều chỉnh lại dữ liệu cho hợp lý.
  + **5. Loại bỏ hoặc thay thế dữ liệu ngoại lai (Outliers)**
  + **Xác định các giá trị ngoại lai** (outliers), những điểm dữ liệu cách xa phần lớn các giá trị còn lại. Đây có thể là các lỗi hoặc giá trị cực đoan.
  + Quyết định xem có nên **loại bỏ, giữ lại, hoặc thay thế** chúng dựa trên bối cảnh dữ liệu và mục tiêu phân tích.
  + **6. Xử lý định dạng dữ liệu không đúng (Data Type Issues)**
  + Đảm bảo rằng các cột dữ liệu có kiểu dữ liệu đúng. Ví dụ: số liệu nên được lưu dưới dạng số (int, float), và ngày tháng nên là kiểu ngày tháng (datetime).
  + Chuyển đổi kiểu dữ liệu nếu cần.
  + **7. Xử lý dữ liệu không nhất quán (Inconsistent Data)**
  + Đảm bảo các giá trị dữ liệu nhất quán. Ví dụ: tên thành phố có thể bị viết khác nhau (Hà Nội, Ha Noi), cần chuẩn hóa lại để tránh nhầm lẫn.
  + Sửa lỗi chính tả hoặc lỗi nhập liệu trong văn bản.
  + **8. Xử lý lỗi mã hóa ký tự (Encoding Issues)**
  + Đối với dữ liệu văn bản, có thể gặp phải các vấn đề về mã hóa ký tự (ví dụ: UTF-8, ASCII) dẫn đến việc hiển thị sai ký tự đặc biệt hoặc ngôn ngữ không phải tiếng Anh. Cần xác định đúng mã hóa và sửa lại.
  + **9. Xử lý dữ liệu phân loại (Categorical Data)**
  + Đối với dữ liệu dạng phân loại (categorical data), cần đảm bảo các nhãn phân loại không bị nhầm lẫn hoặc không nhất quán.
  + Có thể cần **chuyển đổi dữ liệu phân loại thành các biến số** có thể sử dụng được trong các mô hình phân tích (ví dụ: one-hot encoding).
  + **10. Tạo các biến mới (Feature Engineering)**
  + Trong một số trường hợp, việc tạo ra các biến mới dựa trên dữ liệu hiện có sẽ giúp cải thiện chất lượng dữ liệu và mô hình. Ví dụ: từ dữ liệu ngày tháng, bạn có thể tạo thêm biến tháng hoặc năm.
  + **11. Xóa bỏ dữ liệu không cần thiết (Irrelevant Data)**
  + Loại bỏ các cột hoặc hàng không có giá trị cho phân tích (ví dụ: cột chứa thông tin không liên quan đến mục tiêu phân tích).
  + **12. Kiểm tra và khắc phục lỗi logic**
  + Kiểm tra xem dữ liệu có tuân theo các quy tắc logic không. Ví dụ: ngày kết thúc không thể trước ngày bắt đầu, hoặc tổng không thể âm
* Chuẩn hóa dữ liệu
* Biểu diễn dữ liệu ở dạng đồ thị
* Tạo giao diện
  + mục tiêu của ứng dụng là tập trung vào xử lý và phân tích dữ liệu, thì phần mô tả này có thể được lưu trữ dưới dạng tài liệu đi kèm hoặc phần giải thích trong báo cáo cuối cùng thay vì đưa vào giao diện chính.